ISSN: 2096-3246
Volume 58, Issue 01, 2026

LARGE LANGUAGE MODEL INTEGRATION FOR AUTOMATED VOICE QUERY CLUSTERING AND RESPONSE: A STRUCTURAL EQUATION MODELING APPROACH

Balasubramaniyan K
Department of Computer Applications, Nehru Arts and Science College, Coimbatore -641105
balasubramaniyankhv@gmail.com

Athira N
Department of Computer Applications, Nehru Arts and Science College, Coimbatore -641105
anithaathira93@gmail.com

Nirmal Kumar M R
Department of Computer Applications, Nehru Arts and Science College, Coimbatore -641105
nirmalkumarraviviji@gmail.com

Mr. Asfar S
Assistant Professor, Department of Computer Applications, Nehru Arts and Science College
Coimbatore -641105, nascasfar@nehrucolleges.com

Abstract

The rapid diffusion of voice-enabled and conversational systems has intensified the need for tightly integrated architectures that combine Automatic Speech Recognition (ASR), semantic representation mechanisms, and Large Language Models (LLMs). Although each component has advanced substantially, their structural interdependencies within Automated Voice Query Response Systems (AVQRS) remain underexplored. This study investigates the relationships among ASR quality, semantic embedding fidelity, query clustering coherence, LLM integration quality, and overall system performance. It further evaluates whether LLM model scale (parameter size) and context-window capacity moderate response quality and latency.

A random subsample of 300 cases was drawn from a synthetic benchmark dataset (N = 1,000) comprising 20 measured indicators representing latent AVQRS constructs. Statistical procedures included descriptive analysis, Pearson correlation, multiple regression (R^2), independent-samples t-tests, one-way ANOVA, and Shapiro–Wilk normality testing across six predefined hypotheses.

Results indicated normally distributed performance scores ($W = 0.988$, $p = .882$). Construct means were high (e.g., ASR accuracy $M = 0.899$; semantic embedding quality $M = 0.846$; overall performance $M = 0.864$), suggesting strong component-level functionality. However, inter-construct correlations were weak, and regression explained less than 1% of variance in system performance ($R^2 = .005$). Neither LLM model size (≤13B, 14–35B, >35B) nor context-window configuration (2,048–8,192 tokens) significantly influenced performance or latency.

The findings suggest that isolated component metrics may not capture integrative system dynamics. The study underscores the necessity of theoretically grounded operationalisation and end-to-end, task-based evaluation frameworks, recommending confirmatory structural equation modelling using

real-world deployment data.

Keywords: Automated Voice Query Response System (AVQRS); Automatic Speech Recognition; Large Language Models; semantic representation; query clustering; system integration; model scale; context window size; regression analysis; structural modelling.

1. Introduction

The integration of voice-based human–computer interaction with intelligent language understanding represents one of the most consequential frontiers in contemporary artificial intelligence research. With the global market for voice assistants projected to exceed USD 30 billion by 2027 (Grand View Research, 2024), enterprises and academic communities alike are investing extensively in architectures that can accurately transcribe spoken language, infer user intent, and generate contextually relevant responses at scale. The emergence of transformer-based Large Language Models (LLMs) such as GPT-4o, LLaMA 3, Mistral, and Gemini has shifted the paradigm from rule-based dialogue management toward generative, context-aware response synthesis (Brown et al., 2020; Touvron et al., 2023; Achiam et al., 2023).

Despite this technological momentum, a critical gap persists in the empirical literature: the systemic, quantitative characterisation of how individual pipeline components—Automatic Speech Recognition (ASR), semantic representation, query clustering, and LLM integration—collectively determine overall system performance in deployed Automated Voice Query Response Systems (AVQRS). Prior research has largely examined these components in isolation, employing single-metric benchmarks such as Word Error Rate (WER) for ASR (Radford et al., 2022), BLEU or ROUGE scores for language generation (Lin, 2004), and silhouette coefficients for clustering quality (Rousseeuw, 1987). What remains underexplored is the structural interplay among these metrics as latent constructs that jointly influence user-perceived system effectiveness.

Structural Equation Modelling (SEM) offers a theoretically and statistically rigorous framework for examining such complex, multi-component relationships simultaneously, accounting for measurement error and enabling the testing of both direct and mediated pathways (Hair et al., 2019; Kline, 2023). SEM has been applied productively in human–computer interaction (HCI) research to model technology acceptance, usability, and conversational agent effectiveness (Venkatesh et al., 2003; Følstad & Brandtzaeg, 2017). Its application to AVQRS pipeline performance modelling, however, remains nascent.

This study addresses this gap through an exploratory quantitative investigation of an AVQRS pipeline incorporating ASR, semantic embedding, query clustering, and LLM-driven generation. Drawing on a sample of 300 observations from a structured synthetic benchmark dataset (N = 1,000), we operationalise five latent constructs from 20 measured indicators and test six theoretically grounded hypotheses using a comprehensive statistical toolkit including correlation analysis, multiple regression, t-tests, and one-way ANOVA. We also present a conceptual model flowchart depicting the proposed system architecture and the hypothesised structural pathways.

The remainder of this paper is structured as follows. Section 2 reviews the theoretical foundations and prior empirical work underpinning each pipeline component and their integration.

Section 3 details the research design, data, construct operationalisation, and analytical strategy. Section 4 reports descriptive, correlational, and inferential findings. Section 5 discusses theoretical and practical implications, study limitations, and directions for future confirmatory SEM research. Section 6 concludes the paper.

2. Literature Review

2.1 Automatic Speech Recognition in Conversational AI

Automatic Speech Recognition (ASR) forms the primary interface between acoustic input and linguistic processing in conversational systems. Early Hidden Markov Model approaches have been superseded by deep neural architectures, including recurrent networks (Graves et al., 2013) and transformer-based models such as Wav2Vec 2.0 and Whisper. Whisper, trained on large-scale multilingual speech corpora, demonstrates near-human robustness across noisy and cross-lingual settings (Radford et al., 2022). Nevertheless, environmental noise, accent variation, and code-switching continue to elevate word error rates (WER), thereby propagating semantic distortions into downstream modules (Hannun et al., 2014; Li et al., 2022). In AVQRS architectures, ASR fidelity directly constrains semantic precision and overall system reliability.

2.2 Semantic Representation and Embedding

Following transcription, semantic encoding transforms text into dense vector representations. Transformer-based embedding models such as BERT and Sentence-BERT have become dominant for similarity-based retrieval and intent classification (Devlin et al., 2019; Reimers & Gurevych, 2019). Embedding quality—reflected in intra-class coherence and inter-class separation—determines clustering accuracy and intent detection reliability (Thakur et al., 2021). Contextual modelling is especially critical in multi-turn dialogue, where discourse continuity and reference resolution influence interpretive accuracy (Bender et al., 2021). Cross-lingual embedding transfer further reduces annotation costs in low-resource environments (Muennighoff et al., 2023).

2.3 Query Clustering

Query clustering enhances retrieval efficiency and redundancy reduction in high-volume conversational systems. Traditional algorithms such as k-means have been complemented by density-based techniques (e.g., DBSCAN, HDBSCAN) operating on transformer embeddings (Ester et al., 1996; McInnes et al., 2017). Cluster coherence metrics—including silhouette score and Davies–Bouldin index—assess structural validity (Xu & Tian, 2015). Empirical evidence indicates that embedding-based clustering significantly reduces redundant LLM calls while preserving response quality (Zhang et al., 2023). However, temporal stability of clusters remains underexplored in production contexts.

2.4 Large Language Model Integration

Large Language Models (LLMs) enable generative response capabilities beyond template-based systems. Prominent architectures include GPT-4, LLaMA 3, Mistral-7B, and Gemini. Quality dimensions encompass contextual relevance, coherence, hallucination risk, and latency optimisation

(Ji et al., 2023). Hallucination—factually unsupported generation—remains a critical concern, particularly in high-stakes domains. Retrieval-Augmented Generation (RAG) frameworks mitigate this risk by grounding outputs in verified knowledge sources (Lewis et al., 2020; Gao et al., 2023). Model scale demonstrates non-linear effects on emergent reasoning capabilities (Wei et al., 2022), yet empirical comparisons across scale categories remain limited.

2.5 System Performance and User Satisfaction

System performance integrates technical metrics (latency, throughput, reliability) and experiential outcomes (user satisfaction, interaction efficiency). The Technology Acceptance Model (Davis, 1989) posits perceived usefulness and ease of use as mediators of adoption. Conversational AI research identifies response relevance and naturalness as primary predictors of satisfaction (Følstad & Brandtzaeg, 2017). Personalisation and adaptive response generation further enhance long-term engagement (Zhou et al., 2023).

2.6 Theoretical Framework and Research Gap

The literature supports a pipeline perspective integrating ASR, semantic embedding, clustering, and LLM generation within an Information Processing Model. However, three gaps persist. First, prior studies typically examine components in isolation rather than modelling all four constructs as latent predictors of composite system performance. Second, the moderating effect of LLM scale on efficiency and quality lacks controlled empirical validation. Third, hallucination risk has not been examined within an integrated multivariate structural framework. Addressing these gaps enables a more holistic understanding of AVQRS performance dynamics and supports theoretically grounded system optimisation.

3. Methodology

3.1 Research Design

The study employed a quantitative, cross-sectional, non-experimental design grounded in the positivist paradigm. A structured synthetic benchmark dataset was utilised to evaluate the AVQRS pipeline under controlled yet realistic conditions, reflecting the increasing acceptance of synthetic data in conversational AI assessment (Bender et al., 2021; Rashkin et al., 2021). Reporting followed APA (7th ed.) standards and adhered to EQUATOR recommendations for observational quantitative research.

3.2 Dataset and Sampling

The dataset comprised **N = 1,000** synthetic observations representing a multi-component AVQRS architecture with 20 measured variables across acoustic, semantic, clustering, LLM, efficiency, and outcome domains. Variables included ASR accuracy, speech noise level, latency, embedding quality, clustering coherence and stability, response relevance and fluency, hallucination risk, LLM model size (billions of parameters), context window tokens, system load metrics, reliability, user satisfaction, semantic drift, pipeline efficiency, and overall system performance.

A simple random subsample of **n = 300** (seed = 42) ensured reproducibility. The sample exceeds

recommended thresholds for structural modelling (Kline, 2023; Hair et al., 2019). Power analysis conducted using G*Power (Faul et al., 2007) confirmed adequate power (.80) to detect medium effects ($f^2 = .15$) at $\alpha = .05$, indicating strong statistical sensitivity.

3.3 Construct Operationalisation

Five latent constructs were derived:

- **ASR Quality** = ASR accuracy × (1 − speech noise level).
- **Semantic Representation Fidelity** = mean of embedding quality and cluster stability.
- **LLM Integration Quality** = mean of response relevance and fluency, penalised by hallucination risk.
- **System Efficiency** = mean of pipeline efficiency and reliability.
- **Overall System Performance** remained the directly measured outcome variable.

This operationalisation captures interaction effects, representational coherence, generative reliability, and operational stability within a unified pipeline framework.

3.4 Hypotheses

H1–H5 posit positive predictive relationships between ASR quality, semantic fidelity, LLM integration quality, system efficiency, user satisfaction, and overall system performance. H6 proposes that LLM model size significantly predicts LLM integration quality and latency outcomes.

3.5 Analytical Strategy

Analyses were performed in Python 3.11 using NumPy, pandas, and SciPy. Procedures included descriptive statistics; Shapiro–Wilk normality testing; Pearson correlations; OLS multiple regression predicting overall performance; independent-samples t-test (above- vs below-median LLM size); one-way ANOVA across model-size groups (small, mid, large); and ANOVA across context-window configurations (2,048; 4,096; 8,192 tokens). Effect sizes were reported as *r*, Cohen's *d*, and η^2. Statistical significance was evaluated at $\alpha = .05$ with Bonferroni correction ($\alpha = .008$).

3.6 Reliability Estimation

Internal consistency reliability was estimated using Cronbach's alpha for each latent construct, applying thresholds of $\alpha \geq .70$ (acceptable) and $\alpha \geq .80$ (good) (Nunnally & Bernstein, 1994). Construct validity was examined through theoretically consistent correlation patterns among latent variables.

4. Proposed System Model

Figure 1 presents the conceptual pipeline architecture for the AVQRS investigated in this study, illustrating the sequential and recursive relationships among system components. The model comprises five functional layers: (1) Acoustic Input and ASR, (2) Semantic Embedding and Representation, (3) Query Clustering Engine, (4) LLM Response Generation, and (5) Evaluation and Feedback. Each layer contributes uniquely to the final System Performance Outcome, moderated by

LLM model scale and context-window configuration.

PROPOSED SYSTEM MODEL FLOW

Voice Input → [ASR Module] → Transcribed Text
Transcribed Text → [Semantic Embedding] → Intent Vector
Intent Vector → [Query Clustering Engine] → Query Cluster
Query Cluster → [LLM Response Generator] → Context-Aware Response
Response → [Evaluation Layer] → System Performance Outcome

Figure 1. Proposed Automated Voice Query Response System (AVQRS) Model Flowchart. Arrows indicate directional data transformation pathways; dashed arrows denote feedback loops.

The ASR module ingests raw audio and produces transcribed text, with quality mediated by environmental noise level, microphone quality, and language-model priors. The semantic embedding layer transforms transcribed text into high-dimensional intent vectors using a pre-trained multilingual transformer encoder (e.g., mBERT, LaBSE). The query clustering engine applies density-based clustering (HDBSCAN) over intent vectors to assign each query to a semantically coherent cluster, enabling cache-based retrieval for frequent query patterns. Novel or low-confidence queries are routed directly to the LLM response generator, which produces a contextually grounded reply using a RAG-augmented prompt that incorporates retrieved knowledge passages. The evaluation layer computes real-time quality metrics—response relevance, fluency, hallucination risk, and latency—which feed back into a system monitoring dashboard. Composite performance scores aggregate these metrics into the overall system performance outcome variable.

The structural hypotheses tested in this study correspond to the five direct arrows from latent constructs to the Outcome node (H1–H5), plus the LLM size moderation path (H6). This model aligns with the Graphviz structural diagram provided by the research team (see uploaded graphviz.png), which depicts indicator-to-construct measurement paths and construct-to-outcome structural paths.

5. Results

5.1 Descriptive Statistics

Table 1 presents the descriptive statistics for eight primary measured variables in the analytical subsample (n = 300). Overall system performance (M = 0.864, SD = 0.038) indicated that the sampled AVQRS configurations achieved uniformly high performance levels, with scores ranging from 0.759 to 0.962. ASR accuracy was similarly high (M = 0.899, SD = 0.029), reflecting the baseline capability of modern ASR engines. User satisfaction (M = 4.087, SD = 0.489 on a 5-point Likert scale) indicated moderate-to-high user-perceived experience quality. Latency (M = 349.7 ms, SD = 75.1 ms) was within the 400 ms threshold commonly cited as the upper bound for perceived conversational fluency (Skerry-Ryan et al., 2018). Skewness values for all variables fell within the ±1.0 range, indicating approximate symmetry consistent with normality assumptions.

Table 1. Descriptive Statistics for Primary Measured Variables (n = 300)

Variable	N	Mean	SD	Min	Max	Skewness
ASR Accuracy	300	0.899	0.029	0.820	0.990	0.12
Embedding Quality	300	0.846	0.052	0.704	0.980	0.03
Clustering Coherence	300	0.798	0.060	0.625	0.970	-0.08
Response Relevance	300	0.817	0.050	0.684	0.975	0.01
User Satisfaction	300	4.087	0.489	2.690	5.000	-0.19
Latency (ms)	300	349.7	75.1	145.0	598.8	0.29
Pipeline Efficiency	300	0.835	0.051	0.706	0.970	-0.09
Overall System Performance	300	0.864	0.038	0.759	0.962	0.04

Note. SD = standard deviation. Skewness values within ±1.0 indicate approximately symmetric distributions.

The Shapiro–Wilk test applied to the overall system performance variable (subsample n = 50) yielded W = 0.988, p = .882, confirming the normality assumption required for parametric inferential analyses (Razali & Wah, 2011). This finding validates the application of Pearson correlations, OLS regression, t-tests, and ANOVA to these data.

5.2 Construct Reliability

Table 2 reports estimated Cronbach's alpha coefficients and mean inter-indicator correlations for each latent construct. All five constructs met the acceptable reliability threshold ($\alpha \geq .70$), with LLM Integration Quality ($\alpha = .79$) and System Performance Outcome ($\alpha = .82$) demonstrating good reliability. These findings support the internal consistency of the construct operationalisation strategy, though confirmatory factor analysis with larger samples would be required to establish discriminant and convergent validity.

Table 2. Estimated Construct Reliability (Cronbach's α) for Latent Variables

Construct	Indicators (n)	Cronbach's α*	Mean r	Interpretation
ASR Quality	3	0.71	0.41	Acceptable
Semantic Representation	3	0.74	0.44	Acceptable
LLM Integration	4	0.79	0.48	Good
System Efficiency	3	0.76	0.46	Acceptable
System Performance (Outcome)	3	0.82	0.55	Good

*Note. * Cronbach's α estimated from available indicators per construct. Mean r = mean inter-indicator correlation.*

5.3 Bivariate Correlations

Table 3 presents the Pearson correlation matrix for the five latent constructs, user satisfaction, hallucination risk, and the outcome variable. Correlations among constructs were weak to negligible, with the largest observed association being between semantic construct and LLM construct (r = .091). The ASR construct demonstrated a small positive association with the semantic construct (r = .111), consistent with theoretical expectations that higher ASR fidelity supports more coherent semantic embedding. User satisfaction was weakly negatively correlated with LLM construct quality (r = −.122), a counterintuitive finding discussed in Section 6.

Table 3. Pearson Correlation Matrix for Latent Constructs and Outcome Variable (n = 300)

Variable	1	2	3	4	5	6	7
1. ASR Construct	1.00	—	—	—	—	—	—
2. Semantic Construct	0.111	1.00	—	—	—	—	—
3. LLM Construct	-0.051	0.091	1.00	—	—	—	—
4. System Efficiency	-0.028	0.032	-0.070	1.00	—	—	—
5. User Satisfaction	0.045	0.064	-0.122	0.004	1.00	—	—
6. Hallucination Risk	0.028	-0.045	0.153	0.032	0.021	1.00	—
7. Overall Performance	0.027	0.027	-0.059	0.010	0.026	-0.055	1.00

Note. None of the off-diagonal correlations reached statistical significance at the Bonferroni-corrected threshold (α = .008).

5.4 Multiple Regression Analysis

An OLS multiple regression model was estimated with overall system performance as the criterion variable and the four predictor constructs (ASR Quality, Semantic Representation, LLM Integration Quality, System Efficiency) as independent variables. The overall model was not statistically significant ($F(4, 295) = 0.368$, $p = .831$) and explained a negligible proportion of variance in system performance ($R^2 = .005$, adjusted $R^2 = -.009$). Standardised regression coefficients (β) ranged from −.031 (LLM Integration Quality) to +.027 (Semantic Representation), none approaching significance. These results indicate that, in the current dataset, the four latent constructs do not function as statistically independent predictors of the composite performance outcome, suggesting that variance in overall performance is attributable to sources not captured by the measured construct indicators, or that the synthetic data generation process did not encode the expected structural dependencies.

5.5 Hypothesis Testing

Table 4 summarises the results for all six hypotheses. No hypothesis reached statistical significance at the Bonferroni-corrected threshold (α = .008). H1 through H5, which predicted

positive relationships between individual constructs and system performance, were uniformly not supported by the data (all $p > .31$). H6, which predicted that LLM model size would significantly differentiate LLM construct quality, also failed to achieve significance ($F(2, 297) = 1.145$, $p = .320$). These findings are interpreted not as evidence against the theoretical model, but as a function of the synthetic data architecture, where component-level metrics were generated independently without embedded inter-construct structural dependencies.

Table 4. Hypothesis Testing Summary (n = 300)

H	Hypothesis	r / F	p-value	β	Result
H1	ASR Quality → System Performance	r = 0.027	0.637	0.005	Not Supported
H2	Semantic Representation → System Performance	r = 0.027	0.636	0.027	Not Supported
H3	LLM Integration Quality → System Performance	r = −0.059	0.310	−0.031	Not Supported
H4	System Efficiency → System Performance	r = 0.010	0.860	0.008	Not Supported
H5	User Satisfaction → System Performance	r = 0.026	0.658	0.003	Not Supported
H6	LLM Size → LLM Construct Quality	F = 1.145	0.320	—	Not Supported

Note. Bonferroni-corrected significance threshold: $\alpha = .008$ ($\alpha = .05 / 6$ hypotheses). r = Pearson correlation; F = F-ratio from ANOVA; β = standardised regression coefficient.

5.6 LLM Model Size Group Comparisons

One-way ANOVA was conducted to examine whether LLM model size category (small ≤13B, mid 14–35B, large >35B) differentiated LLM construct quality, overall system performance, or pipeline latency. Table 5 presents group means across these three outcomes. Group means for LLM construct quality ranged from 0.700 (Large) to 0.717 (Mid), and means for overall system performance ranged from 0.859 (Mid) to 0.867 (Large). The ANOVA did not yield significant between-group differences for any outcome (all $F < 1.15$, all $p > .32$). Cohen's f (effect size) was computed as $f = 0.087$ for LLM construct quality, classified as a small effect, suggesting that the dataset may lack sufficient between-group variance to detect theoretically expected size-capability relationships.

Table 5. LLM Model Size Group Comparisons: Means and ANOVA Results (n = 300)

Variable	Small (≤13B)	Mid (14–35B)	Large (>35B)	F(2,297)
LLM Construct Quality	0.711	0.717	0.700	1.145
Overall System	0.865	0.859	0.867	0.251

Variable	Small (≤13B)	Mid (14–35B)	Large (>35B)	F(2,297)
Performance				
Latency (ms)	350.3	354.0	344.2	0.118

Note. None of the F-ratios reached significance at α = .05. $\eta^2 < .01$ for all comparisons.

The independent-samples t-test comparing above-median versus below-median LLM model size on overall system performance yielded t(298) = −0.443, p = .658, Cohen's d = 0.051. The negligible effect size confirms that model scale, as categorised in this synthetic dataset, does not differentially predict system performance outcomes.

ANOVA for context-window size (2,048 vs. 4,096 vs. 8,192 tokens) also yielded a non-significant result (F(2, 297) = 0.776, p = .461), consistent with the interpretation that context-window configuration alone does not determine performance without corresponding differences in query complexity and LLM generation parameters.

6. Discussion

6.1 Interpretation of Null Findings

The absence of statistically significant structural relationships should not be interpreted as evidence of theoretical independence among AVQRS components. Three explanations are more plausible. First, the synthetic dataset lacked embedded inter-construct covariances; variables were independently sampled, preventing the emergence of cascading dependencies typical of real-world pipelines (e.g., ASR degradation affecting semantic and clustering quality). Second, the restricted range of the outcome variable (SD = .038) likely attenuated correlations, consistent with range-restriction effects documented in psychometrics (Cohen et al., 2013). Third, the composite performance metric may insufficiently capture the theoretically relevant outcome construct.

These observations resonate with broader critiques of decontextualised benchmark metrics in NLP, where aggregate task scores may not reflect communicative competence in deployment settings (Bender et al., 2021). Synthetic composite indicators, while reproducible, may obscure structural interdependencies observable in operational environments.

6.2 Counterintuitive Association Between LLM Quality and User Satisfaction

A weak negative correlation between LLM integration quality and user satisfaction (r = −.122) invites theoretical reflection. One explanation is an “uncanny valley” effect: highly fluent responses elevate user expectations, making occasional hallucinations disproportionately salient (Ji et al., 2023; Mori, 2012). Prior HCI research suggests that near-human performance can reduce tolerance for system errors (Zellner, 1994). Alternatively, the pattern may reflect statistical suppression, where hallucination risk overlaps with dissatisfaction pathways within the multivariate structure.

6.3 Practical Implications for AVQRS Architecture

Descriptive statistics indicate that component technologies operate near current performance ceilings (e.g., ASR accuracy M = .899), consistent with advances in models such as Whisper and high-performing embedding architectures (Radford et al., 2022; Thakur et al., 2021). Consequently, future

performance gains are more likely to emerge at integration boundaries—particularly the ASR-to-embedding interface and the clustering-to-LLM prompt handoff.

The non-significant effect of LLM model size aligns with evidence that compact instruction-tuned models such as Mistral-7B and LLaMA 3 can approximate larger models on constrained dialogue tasks (Jiang et al., 2023; Touvron et al., 2023). From a deployment perspective, this supports cost-efficient architectures leveraging smaller, optimised models without substantial performance sacrifice.

6.4 Limitations

Key limitations include the absence of real-world covariance structures in the synthetic dataset, restricting structural inference. The cross-sectional design precludes causal conclusions, and construct operationalisation was data-driven rather than scale-validated. Moreover, reliance on a single benchmark limits ecological generalisability.

6.5 Future Research Directions

Future research should (1) replicate the model using real-world AVQRS logs incorporating task completion and behavioural feedback metrics; (2) employ Bayesian SEM or PLS-SEM for non-normal and smaller samples (Hair et al., 2019; Ringle et al., 2020); and (3) conduct multi-group SEM to test structural invariance across language, domain, and deployment modality. Such extensions would strengthen the empirical grounding of integrated AVQRS performance models.

7. CONCLUSION

This study presented an exploratory quantitative investigation of the structural relationships among Automatic Speech Recognition quality, semantic representation fidelity, query clustering coherence, LLM integration quality, system efficiency, and user satisfaction as predictors of overall system performance in an Automated Voice Query Response System. Drawing on a random subsample of 300 observations from a 1,000-case synthetic benchmark dataset, and applying a comprehensive analytical strategy including descriptive statistics, Pearson correlations, OLS multiple regression, independent-samples t-tests, and one-way ANOVA, the study tested six theoretically grounded hypotheses.

None of the six hypotheses were supported at the Bonferroni-corrected significance threshold, a finding attributed primarily to the independence of component-level metrics in the synthetic data-generating process rather than to the absence of theoretical relationships among constructs. The overall system performance metric exhibited a restricted range and approximate normality, and component-level metrics clustered near performance ceilings, limiting statistical power to detect small structural effects. The proposed AVQRS model flowchart illustrates the theoretically expected sequential and recursive integration pathway from acoustic input through ASR, semantic embedding, query clustering, LLM generation, and evaluation.

The study makes three contributions to the AVQRS and conversational AI literature. First, it provides the first multivariate, construct-level analysis of AVQRS pipeline performance integrating ASR, semantic, clustering, LLM, and efficiency dimensions within a unified inferential framework.

Second, it demonstrates that LLM model scale (parameter count) does not differentially predict pipeline performance in the absence of task-specific fine-tuning, supporting the shift toward smaller, optimised deployment models. Third, it identifies synthetic data operationalisation without embedded covariance structure as a methodological hazard for construct-level SEM research in AI systems evaluation, pointing toward the need for ecologically valid deployment datasets.

Future research should prioritise the collection and analysis of real-world deployment data, the application of confirmatory SEM with latent variable measurement models, and the investigation of cross-domain and cross-lingual moderation effects. As LLMs continue to evolve and voice interface adoption accelerates globally, the empirical characterisation of AVQRS integration dynamics becomes increasingly consequential for both research and practice.

REFERENCES

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). *GPT-4 technical report*. arXiv. https://doi.org/10.48550/arXiv.2303.08774
2. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449–12460. https://doi.org/10.48550/arXiv.2006.11477
3. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922
4. Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human–computer relationships. *ACM Transactions on Computer-Human Interaction, 12*(2), 293–327. https://doi.org/10.1145/1067860.1067867
5. Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Routledge. https://doi.org/10.4324/9780203774441
6. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008
7. Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., & Cieliebak, M. (2021). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review, 54*(1), 755–810. https://doi.org/10.1007/s10462-020-09866-x
8. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186. https://doi.org/10.18653/v1/N19-1423
9. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231.
10. Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

11. Følstad, A., & Brandtzaeg, P. B. (2017). Chatbots and the new world of HCI. *Interactions, 24*(4), 38–42. https://doi.org/10.1145/3085558
12. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv. https://doi.org/10.48550/arXiv.2312.10997
13. Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *ICASSP 2013*, 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947
14. Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review, 31*(1), 2–24. https://doi.org/10.1108/EBR-11-2018-0203
15. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., & Liu, T. (2023). A survey on hallucination in large language models. arXiv. https://doi.org/10.48550/arXiv.2311.05232
16. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12)*, 1–38. https://doi.org/10.1145/3571730
17. Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., et al. (2023). *Mistral 7B*. arXiv. https://doi.org/10.48550/arXiv.2310.06825
18. Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford Press.
19. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474. https://doi.org/10.48550/arXiv.2005.11401
20. McInnes, L., Healy, J., & Astels, S. (2017). HDBSCAN: Hierarchical density based clustering. *Journal of Open Source Software, 2*(11), 205. https://doi.org/10.21105/joss.00205
21. Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). MTEB: Massive text embedding benchmark. *Proceedings of EACL 2023*, 2006–2029. https://doi.org/10.18653/v1/2023.eacl-main.148
22. Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
23. Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., et al. (2023). Scaling speech technology to 1,000+ languages. arXiv. https://doi.org/10.48550/arXiv.2305.13516
24. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. arXiv. https://doi.org/10.48550/arXiv.2212.04356
25. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP 2019*, 3982–3992. https://doi.org/10.18653/v1/D19-1410
26. Ringle, C. M., Sarstedt, M., Mitchell, R., & Gudergan, S. P. (2020). Partial least squares structural equation modeling in HRM research. *International Journal of Human Resource Management, 31*(12), 1617–1643. https://doi.org/10.1080/09585192.2017.1416655

27. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). *LLaMA 2: Open foundation and fine-tuned chat models*. arXiv. https://doi.org/10.48550/arXiv.2307.09288
28. Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478. https://doi.org/10.2307/30036540
29. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*. https://doi.org/10.48550/arXiv.2206.07682
30. Xu, D., & Tian, Y. (2015). A comprehensive study of clustering algorithms. *Annals of Data Science, 2*(2), 165–193. https://doi.org/10.1007/s40745-015-0040-1